Poster II-44

A Unified Biological Signal and Image Data Format for Compression, Integration, and Privacy Protection

Sun, Mingui*[*1], Shi, Yunqing[2], Liu, Qiang[1], Sclabassi, Robert J.[1]

[1]Laboratory for Computational Neuroscience, Department of Neurosurgery, Electrical Engineering, and Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA; [2]Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA

In current medical information systems, images and waveforms are represented digitally in large-sized data files. Large file size often causes problems in efficient database management and rapid network transmission. In addition, for the purpose of diagnosis, primary medical data must be utilized in association with ancillary data, such as demographic information of patient, data acquisition parameters, and comments and notes at specific points within the data. The primary and ancillary data are often bundled at the time of data access. This method has several drawbacks: 1) it depends on the accessibility of each information source, 2) it involves a risk of data mismatch that may cause serious consequences, and 3) it requires special mechanisms to handle temporally or spatially synchronized data. An alternative method is to place ancillary data in a header of the primary data and time-stamp (or space-stamp) each data record. This method lacks flexibility in cases where multiple ancillary data with different lengths and time clocks (or spatial scales) are integrated. Another problem is the lack of flexibility in access control. A number of federal laws and regulations, e.g., the recent Health Insurance Portability and Accountability Act (HIPPA), restrict the direct use of patient data for the purpose of research. In certain cases, 28 types of identifiers of human subjects must be removed. Manually removing these identifiers and establishing linkage codes by a "honest broker" are extremely time consuming.

We present a unified approach to data compression, integration, and information protection by mixing medical waveforms and images with encrypted patient identifiers and un-encrypted ancillary information, such as acquisition parameters and diagnostic notes. We take advantage of nonstationarity in medical waveforms and images by effectively changing a sampling grid according to the local smoothness of the data. Redundant samples (or pixels) are eliminated which are replaced by ancillary data samples. In this way, closely related information is bundled into a single file. In order to achieve data security, ancillary data are classified into sensitive and non-sensitive categories. For sensitive data, a special-purpose encryption algorithm is used which not only denies access to this type of data, but also automatically provides unique linkage codes accessible to the researchers. Conversion of these codes to the original data can only be performed by authorized individuals with security keys. For non-sensitive information, the ancillary data are accessible, but are watermarked into the waveforms and images. We provide an efficient and versatile technique of using a status string to label the integrated data points, and compress this string by grouping its elements and applying the Huffman and run-length coding algorithms.

Our unified data compression, integration, and privacy protection method has several advantages over the existing methods: 1) it achieves information embedding and data compression at the same time; 2) it is very flexible allowing the user to insert essentially any forms of digitized information; 3) the data size is always non-expansive and automatically adjustable with respect to the size of the embedded information; 4) it inserts information at desired locations and chopping the data does not change the relevance of the embedded information; and 5) it provides a novel security mechanisms to facilitate the compliance with the federal regulations.

This work was supported in part by NIH grants NS38494 and EB002309, and ARMY grant CECOM DAAB07-01-D-G-001.